CLASSICAL AND BAYES-P* SUBSET SELECTION PROCEDURES FOR DOUBLE EXPONENTIAL POPULATIONS*

by

Shanti S. Gupta
Department of Statistics
Purdue University
West Lafayette, IN, USA

and

Yuning Liao
Department of Statistics
Purdue University
West Lafayette, IN, USA

Technical Report # 90-25C

Department of Statistics
Purdue University

May 1990

*Research supported in part by the Office of Naval Research Contract N00014-88-K-0170 and NSF Grant DMS-8702620 at Purdue University.

△

CLASSICAL and BAYES-P^* SUBSET SELECTION PROCEDURES for DOUBLE EXPONENTIAL POPULATIONS*

Shanti S. Gupta and Yuning Liao

Department of Statistics, Purdue University, West Lafayette, Indiana 47907, U.S.A.

Abstract

The exact distribution of the sample mean from a double exponential(Laplace) model is derived. A *classical subset selection procedure* based on the sample mean for selecting the population associated with the largest location parameter of k double exponential(Laplace) distributions is studied. For the case when a non-informative prior is introduced into the problem, the relation between the classical Maximum-Type Procedure Rule $R^{\max}$ and the so-called Bayes-P^* subset selection procedure rule is studied. An improved bound for the guarantee probability of a correct selection for the classical subset selection rule $R^{\max}$ that relates the rule $R^{\max}$ to the selected subset size (notice that the subset selection rule $R^{\max}$ may select all the populations) is studied and some improved rules of the type $R^{\max}$ are provided.

1 Introduction

Suppose we have k double exponential populations $\Pi_1, \Pi_2, \cdots, \Pi_k$, where each Π_i is characterized by the location parameter θ_i, $i = 1, 2, \cdots, k$. The parameters $\theta_1, \theta_2, \cdots, \theta_k$ are assumed to be unknown. Let X_i be the observable random variable from Π_i with probability density function

$$f(x; \theta_i, \sigma) = \frac{\sigma}{2} exp\{-\frac{|x - \theta_i|}{\sigma}\}, \quad -\infty < x, \theta_i < \infty, \ \sigma > 0, \tag{1}$$

*Research supported in part by the Office of Naval Research Contract N00014-88-K-0170 and NSF Grant DMS-8702620 at Purdue University.

where σ, is a common known value for all $i = 1, 2, \cdots, k$, so that without loss of generality, we can assume that $\sigma = 1$. The ranked parameters are denoted by $\theta_{[1]} \leq \theta_{[2]} \leq \cdots \leq \theta_{[k]}$, and it is assumed that the correct pairing of the ordered $\theta_{[i]}$'s and the unordered θ_i's is unknown.

In this paper, we are mainly interested in the subset selection procedures. First, we assume that there is no prior information about the parameters. Then we study the case where θ_i's are independently distributed, and each θ_i has a non-informative prior.

2 Distribution of the Sample Mean

In connection with the selection procedures based on the sample means, we first derive the distribution of the sample mean.

Let X_{ij} be a random sample from ith population $i = 1, 2, \cdots, k, \quad j = 1, 2, \cdots, n$, i.e.

$$X_{ij} \sim f(x|\theta_i, 1) = \frac{1}{2} exp\{-|x - \theta_i|\}.$$

Hence

$$U_{ij} = X_{ij} - \theta_i \sim f(x|0, 1) = \frac{1}{2} exp\{-|x|\}. \tag{2}$$

From the characteristic function of $U_i = \sum U_{ij}$, we can derive the following lemma

Lemma 2.1(Weida (1935)) Suppose $U_i = \sum_{j=1}^{n} U_{ij}$, where U_{ij} has density (2), then the density function of U_i is given by following formula

$$p(u) = \frac{1}{2\pi} 2\pi i \frac{(-1)^{n-1}}{(n-1)!} \frac{1}{i^n} \frac{d^{n-1}}{dt^{n-1}} \{\frac{e^{-itu}}{(1+it)^n}\}|_{t=i^{-1}}, \tag{3}$$

where $u > 0$ and

$$p(u) = p(-u) \quad \text{for} \quad u \leq 0. \qquad \square$$

Let $s = -it$, then (3) becomes

$$\begin{aligned} p(u) &= \frac{1}{(n-1)!} \frac{d^{n-1}}{ds^{n-1}} \{\frac{e^{su}}{(1-s)^n}\}|_{s=-1} \\ &= \frac{e^{-u}}{2^n (n-1)!} \sum_{j=1}^{n} c_{n,n-j} u^{n-j} \\ &= e^{-u} \sum_{j=1}^{n} \frac{c_{n,n-j}}{c_n} u^{n-j}, \end{aligned} \tag{4}$$

where

$$\begin{aligned} c_n &= 2^n(n-1)!, \\ c_{n,n-j} &= \frac{(n+j-2)!}{(j-1)!(n-j)!2^{j-1}}, \quad j = 1,2,\cdots,n. \end{aligned} \tag{5}$$

Therefore the density function of $\overline{X}_i = \sum_{j=1}^{n} X_{ij}/n$ is

$$f_n(|x-\theta_i|) = e^{-n|x-\theta_i|} \sum_{j=1}^{n} \alpha_{n,n-j} |x-\theta_i|^{n-j}, \quad -\infty < x < \infty, \tag{6}$$

where $\alpha_{n,n-j} = n^{n-j+1} c_{n,n-j}/c_n, \quad j = 1,2,\cdots,n.$

To obtain the coefficients $\{c_{n,i}\}$, $i = 0,1,2,\cdots,n-1$, $n = 2,3,\cdots$, it is helpful to rewrite the formula (5) as

$$c_{n,i} = \frac{(2n-i-2)!}{(n-i-1)!i!2^{n-i-1}}. \tag{7}$$

Note that

$$c_{n,n-1} = 1, \quad c_{n,i} = \frac{(2n-i-2)(2n-i-3)}{2(n-i-1)} c_{n-1,i}. \tag{8}$$

In particular

$$c_{n,0} = c_{n,1}, \quad c_{n,1} = (2n-3)c_{n-1,1}.$$

In Table 1, we have provided the values of $\{c_n\}$ and $\{c_{n,i}\}$ for $n = 2(1)10$; $i = 1(1)n-1$.

To find the cdf of $\overline{X}_i$, let us first find the cdf of U_i. Integrating the density function (4) of U_i, we have

$$\begin{aligned} P(u) &= \int_{-\infty}^{u} p(t)dt \qquad (u > 0) \\ &= 1 - e^u \sum_{j=1}^{n} \frac{a_{n,n-j}}{c_n} u^{n-j}, \end{aligned} \tag{9}$$

where $\{a_{n,n-j}\}$ satisfy:

$$a_{n,n-j} = c_{n,n-j} + (n-j+1)a_{n,n-j+1} \quad j = 1,2,\cdots,n, \quad a_{n,n} = 0. \tag{10}$$

Again we have

$$a_{n,n-1} = 1, \quad a_{n,n-2} = (n-1)(n+2)/2.$$

Hence the cdf of $\overline{X}_i$ is given by,

$$F_n(x|\theta_i) = \begin{cases} e^{-n|x-\theta_i|} \sum_{j=1}^{n} \frac{a_{n,n-j} n^{n-j}}{c_n} |x-\theta_i|^{n-j}, & x < \theta_i \\ 1 - e^{-n|x-\theta_i|} \sum_{j=1}^{n} \frac{a_{n,n-j} n^{n-j}}{c_n} |x-\theta_i|^{n-j}, & \text{otherwise.} \end{cases} \tag{11}$$

In Table 2, we provide the values of $\{c_n\}$ and $\{a_{n,i}\}$ for $n = 2(1)10$; $i = 1(1)n-1$.

Example: If we want to obtain the density and the cumulative distribution function of the sample mean of size n=4 from a double exponential model, checking the column $n = 4$ from both Table 1 and Table 2, we can easily see that

$$f_4(|x-\theta_i|) = \frac{1}{96} e^{-4|x-\theta_i|}(4^4|x-\theta_i|^3 + 6 \times 4^3|x-\theta_i|^2 + 15 \times 4^2|x-\theta_i| + 15 \times 4),$$

and

$$F_4(x|\theta_i) = \begin{cases} \frac{1}{96} e^{-4|x-\theta_i|}(4^3|x-\theta_i|^3 + 9 \times 4^2|x-\theta_i|^2 + 33 \times 4|x-\theta_i| + 48), & x < \theta_i, \\ 1 - \frac{1}{96} e^{-4|x-\theta_i|}(4^3|x-\theta_i|^3 + 9 \times 4^2|x-\theta_i|^2 + 33 \times 4|x-\theta_i| + 48), & \text{otherwise.} \end{cases}$$

To compare the percentage points of the sample mean and the sample median, let

$$Z_n = \frac{\overline{X}_n - \theta}{\sigma} \quad \text{and} \quad Z_n^* = \frac{X_{(\frac{n}{2})} - \theta}{\sigma}.$$

Since the cdf of Z_n^* for odd number n is much easier to derive (see Gupta and Leong 1979), we will only provide the comparison of the percentage points of Z_n and Z_n^* for $n = 3, 5, \cdots, 21$(Table 3). The percentage points for the distribution of the sample mean Z_n when $n = 2, 4, \cdots, 20$ are provided in a separate table(Table 4).

3 Using the Sample Mean to Select the Largest Location Parameter

If we assume that no prior information about the parameter $\underline{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)$ is available, then we usually will use either the classical subset selection approach or the indifference zone formulation in our ranking and selection problem. In the following, we only study the subset selection approach.

(A) Formulation of the Problem: The the classical Maximum-Type Approach for any location type problem have been well studied, so we would not give too many details, but simply state some interesting results without any proof.

For selecting the population associated with the largest location parameter with a correct selection(CS) probability at least $P^*(1/k < P^* < 1)$ from k double exponential populations, where we have a sample mean $\overline{X}_i$ of size n from each Π_i $i = 1, 2, \cdots, k$, the Classical Maximum-Type Subset Selection Rule ($R^{\max}$) proposed by Gupta(1956) is defined as follows:

$$R^{\max}: \quad \text{Select } \Pi_i, \quad \text{iff: } \overline{X}_i \geq \max_{1 \leq j \leq k} \overline{X}_j - d/\sqrt{n} \quad \text{for some } d(> 0),$$

where $d(\geq 0)$ is the smallest value satisfying:

$$\int_{-\infty}^{\infty} F_n^{k-1}(u + d/\sqrt{n}) f_n(u) du \geq P^*.$$

The usual condition of $P(CS|R^{\max}) \geq P^*$ is guaranteed by the following theorem:

Theorem 3.1

$$\inf_{\underline{\theta} \in \Omega} P_{\underline{\theta}}(CS|R^{\max}) = \inf_{\underline{\theta} \in \Omega_0} P_{\underline{\theta}}(CS|R^{\max}) = \int_{-\infty}^{\infty} F_n^{k-1}(u + d/\sqrt{n}) f_n(u) du,$$

where $\Omega \supset \Omega_0 = \{\underline{\theta} : \theta_1 = \theta_2 = \cdots = \theta_k, \ -\infty < \theta_i < \infty, \ i = 1, 2, \cdots, k\}$.

(B) Table of Necessary Constants For $R^{\max}$: for given k, n, and some particular values of P^*, the constants $d/\sqrt{n} = d(k, n, P^*)$ which satisfy

$$P^* = \int_{-\infty}^{\infty} F_n^{k-1}(u + d/\sqrt{n}) f_n(u) du,$$

are given in Table 5.

(C) Asymptotic Results for the Procedure $R^{\max}$: For large n, we can certainly use the normal distribution to approximate the infimum of $P_{\underline{\theta}}(CS|R^{\max})$. Since

$$\inf_{\underline{\theta} \in \Omega} P_{\underline{\theta}}(CS|R^{\max}) = \inf_{\underline{\theta} \in \Omega_0} P_{\underline{\theta}}(CS|R^{\max}),$$

it suffices to consider the case where $\underline{\theta} \in \Omega_0$, now we have

$$\frac{\overline{X}_n - \theta}{\sigma_n} \longrightarrow N(0, 1),$$

where $\sigma_n^2 = 2/n$, so the probability of the following event

$$\overline{X}_k \geq \max_{1 \leq j \leq k} \overline{X}_j - d/\sqrt{n},$$

is, asymptotically, the same as that of

$$Z_k \geq \max_{1 \leq j \leq k} Z_j - d/\sqrt{2},$$

where $Z_j,\ j = 1, 2, \cdots, k$ are i.i.d. standard normal variables, thus

$$\begin{aligned} \inf_{\underline{\theta} \in \Omega} P_{\underline{\theta}}(CS|R^{\max}) &\simeq P_{\underline{\theta}}(Z_k \geq \max_{1 \leq j \leq k} Z_j - d/\sqrt{2}) \\ &= \int_{-\infty}^{\infty} \Phi^{k-1}(u + d/\sqrt{2}) d\Phi(u). \end{aligned} \tag{12}$$

On the other hand, if we use the sample median in the selection procedure, we will have, asymptotically,

$$\begin{aligned} \inf_{\underline{\theta} \in \Omega} P_{\underline{\theta}}(CS|R^{\max}_{median}) &\simeq P_{\underline{\theta}}(Z_k \geq \max_{1 \leq j \leq k} Z_j - d_{median}) \\ &= \int_{-\infty}^{\infty} \Phi^{k-1}(u + d_{median}) d\Phi(u). \end{aligned}$$

Thus, in order to have the same probability of a correct selection for both selection rules based on the different statistics, we must have, for large n,

$$d \simeq \sqrt{2} d_{median}. \tag{13}$$

(D) Sensitivity of the Assumption of Double Exponential: Suppose we have k populations $\Pi_1, \Pi_2, \cdots, \Pi_k$, where Π_i is characterized by a location parameter θ_i. If we do not know whether these k populations have normal, logistic, or double exponential distributions, then selecting the population associated with the largest location parameter becomes a problem, because the real distribution of the populations is unknown. We will show that the double exponential distribution model provides a safeguard as explained below.

If the sample size n is large, we know that the infimum of $P_{\underline{\theta}}(CS|R^{\max})$ for the double exponential populations is approximately given by (12). On the other hand, for the normal means problem, we have

$$\inf_{\underline{\theta} \in \Omega} P_{\underline{\theta}}(CS|R^{\max}_N) = \int_{-\infty}^{\infty} \Phi^{k-1}(u + d_N) d\Phi(u),$$

because

$$\begin{aligned} Z_k \geq \max_{1 \leq j \leq k} Z_j - d_N/\sqrt{n} &\iff \\ \sqrt{n}(\overline{Z}_k - \theta) &\geq \max_{1 \leq j \leq k} \sqrt{n}(\overline{Z}_j - \theta) - d_N, \end{aligned}$$

and $\sqrt{n}(\overline{Z}_j - \theta) \sim N(0,1)$. Similarly, for the logistic distribution model, we have

$$\inf_{\underline{\theta} \in \Omega} P_{\underline{\theta}}(CS|R_L^{\max}) \simeq \int_{-\infty}^{\infty} \Phi^{k-1}(u + d_L) d\Phi(u),$$

therefore,

$$d \simeq \sqrt{2} d_N \simeq \sqrt{2} d_L.$$

It is clear from the above that the d-values for the double exponential provide conservative bounds for the other two models, if n is large.

When n is small, for instance, for $n = 10,\ k = 10$, we have the following:

	P^*-value			
	0.75	0.90	0.95	0.99
d	3.1971	4.2510	4.9063	6.1968
d_L	2.2639	2.9925	3.4390	4.3029
d_N	2.2637	2.9829	3.4182	4.2456

d_N-value excerpted from Bechhofer(1954)

d_L-value excerpted from Han(1987 Ph.D. Thesis)

From this we again see that the d-values for the double exponential provide conservative bounds for the normal and logistic models for the problem of selecting the unknown location parameter.

4 Selection Using a Non-informative Prior

In the Classical Maximum-Type Subset Selection Procedure, it is easy to notice that the selected subset size $|s|$ is a random variable which is not fixed in advance.

In general, for any location or scale parameter situation, Gupta(1965) proved that:

(1) The procedure of the above type is monotone, and

(2) If the distribution $F(x, \theta)$ possesses a density $f(x, \theta)$ having a monotone likelihood ratio (MLR) in x, then $E(|s|)$ is maximized when $\theta_1 = \theta_2 = \cdots = \theta_k$ and the maximum is kP^*.

So, in the worst case, the expected proportion in the selected subset is equal to P^*. Furthermore, it may select populations such that, depending on the unknown parameter $\underline{\theta}$, we may get an actual $P(CS)$ much larger than P^*.

In this section, we will regard the likelihood function of θ_i as the distribution of θ_i given $\underline{X}$. It is the same as saying that based on the distribution of a statistic (in our case it is the sample mean and the sample median), we assume that, independently, each Θ_i has a non-informative prior, $i = 1, 2, \cdots, k$.

4.1 Bayes Selection Procedure

In the following, we will consider a more general case, we assume

$$X_i \sim f(|x - \theta_i|),$$

i.e. the density of X_i given $\Theta_i = \theta_i$ is symmetric about θ_i(for the case where $f(.)$ is not symmetrical, we have obtained some results which will be available later), and

$$\Theta_i \sim \Pi(\theta) = 1, \quad i = 1, 2, \cdots, k.$$

Now, we will make decisions based on the posterior distributions of $\underline{\Theta}|\underline{X}$.
From a Bayes perspective, in order to select the population associated with the largest parameter $\theta_{[k]}$ with a guaranteed posterior probability of a correct selection to be at least $P^*(1/k < P^* < 1)$ (the so-called PP^*-condition, see Gupta and Yang(1985)), we should consider the following events

$$\mathcal{A}_i = \{\theta_i \text{ is the largest } |\underline{X} = \underline{x}\}, \quad i = 1, 2, \cdots, k.$$

Now, using the non-informative prior, we have

$$\Theta_i|\underline{X} = \underline{x} \sim f(|x_i - \theta_i|), \quad i = 1, 2, \cdots, k.$$

Let $p_i(\underline{x})$ be the probability of event $\mathcal{A}_i$, then

$$\begin{aligned} p_i(\underline{x}) &= P(\theta_i \text{ is the largest } |\underline{x}) \\ &= P(\theta_i > \theta_j, \forall j,\ j \neq i\ |\underline{x}) \\ &= P(\theta_i - x_i > \theta_j - x_j - (x_i - x_j), \forall j,\ j \neq i\ |\underline{x}) \\ &= \int_{-\infty}^{\infty} \prod_{j \neq i} F(u + (x_i - x_j)) f(u) du, \end{aligned}$$

where $F(.)$ is the cdf of $f(.)$.

Lemma 4.1: (1) The posterior probability $p_i(\underline{x})$ depends only on the differences $x_i - x_j, \quad i, j = 1, 2, \cdots, k, \ j \neq i$,

(2) $p_i(\underline{x})$ is non-increasing in $x_j, \ j \neq i$, keeping other components of $\underline{x}$ fixed and it is non-decreasing in x_i, keeping other components of $\underline{x}$ fixed.

(3) $p_i(\underline{x}) \geq p_j(\underline{x})$ if and only if $x_i \geq x_j$.

Proof: The proof is straightforward and hence omitted. □

Theorem 4.1: For any subset S of the whole populations $\Pi_1, \Pi_2, \cdots, \Pi_k$, let $PP(CS|S, \underline{x})$ denote the posterior probability of a correct selection for the subset S(i.e. the subset S contains the best population) based on a random sample $\underline{x}$, then

(1) $PP(CS|S, \underline{x})$ is non-increasing in x_j, if $j \notin S$, keeping other components of $\underline{x}$ fixed, and

(2) $PP(CS|S, \underline{x})$ is non-decreasing in x_i, if $i \in S$, keeping other components of $\underline{x}$ fixed.

Proof: Since

$$\begin{aligned} PP(CS|S, \underline{x}) &= \sum_{i \in S} p_i(\underline{x}) \\ &= 1 - \sum_{i \notin S} p_i(\underline{x}). \end{aligned} \tag{14}$$

Now, $p_i(\underline{x})$ is non-increasing in x_j, if $j \notin S$ for all $i \in S$, so $PP(CS|S, \underline{x})$ is non-increasing in $x_j, \ j \notin S$ by first part of equation (14).

On the other hand, the second part of equation (14) and the fact that $p_j(\underline{x})$ is non-increasing in x_i, if $i \in S$ for all $j \notin S$ imply that $PP(CS|S, \underline{x})$ is non-decreasing in $x_i, \ i \in S$. □

From the Bayesian analysis, we know that the Bayes Decision Rule (R^B) will select the t populations which associated with the t largest values of $p_i(\underline{x})$ values (i.e. the Bayes set $s^B = \{\Pi_{[k]}, \cdots, \Pi_{[k-t+1]}\}$), where the integer $t(\geq 1)$ satisfies

$$\sum_{m=k-t+1}^{k} p_{[m]}(\underline{x}) \geq P^*,$$

and

$$\sum_{m=k-t+2}^{k} p_{[m]}(\underline{x}) < P^*,$$

where $p_{[1]}(\underline{x}) \leq p_{[2]}(\underline{x}) \leq \cdots \leq p_{[k]}(\underline{x})$ are the ordered values of $p_i(\underline{x})$'s, and s^B is the subset selected by the Bayes selection rule R^B.

4.2 A Lower Bound on the $PP(CS)$ for the Subset Selection Rule $R^{\max}$

Under the Maximum-Type Subset Selection Rule $R^{\max}$ defined in the previous section, we know that the larger the value x_i is, the larger the chance that the corresponding population Π_i will be selected.

Under the rule $R^{\max}$, we will pick the population Π_i if its x_i value is larger than $x_{[k]} - d$, and reject Π_i if $x_i < x_{[k]} - d$. Thus the following observations:

Observations: For the Maximum-Type Subset Selection Rule $R^{\max}$, we know at least the following two facts

(1) $R^{\max}$ will always pick population $\Pi_{[k]}$, i.e. the population associated with the largest value $x_{[k]}$,

(2) All Π_i not being selected by $R^{\max}$ must has its x_i value less than $x_{[k]} - d$.

Theorem 4.2: If the subset selection rule $R^{\max}$ selects i populations(i.e. select population $\Pi_{[k]}, \cdots, \Pi_{[k-i+1]}$, where $\Pi_{[j]}$ is the population associated with the jth largest value $x_{[j]}$), under the classical selection procedure, then

$$\begin{aligned} PP(CS|R^{\max}, \underline{x}) &\geq PP(CS|R^{\max}, \underline{x} \in \mathcal{X}_0) \\ &= P^* + \frac{i-1}{k-1}(1-P^*), \end{aligned} \tag{15}$$

where $\mathcal{X}_0 = \{\underline{x} : \ x_{[k]} - d = x_{[k-1]} = \cdots = x_{[1]} \}$.

Remark 4.2: A similar result for the normal model has been given in Gupta and Yang (1985). Here, we will give a probabilistic proof of the above theorem.

Proof: The first part of the inequality [i.e. $PP(CS|R^{\max}, \underline{x}) \geq PP(CS|R^{\max}, \underline{x} \in \mathcal{X}_0)$] follows from the above observations and Theorem 4.1.

When $\underline{x} \in \mathcal{X}_0$, we know that $p_{[k]}(\underline{x}) > p_{[k-1]}(\underline{x}) = \cdots = p_{[1]}(\underline{x})$, and

$$\begin{aligned} p_{[k]}(\underline{x}) &= \int_{-\infty}^{\infty} \prod_{j \neq k} F(u + (x_{[k]} - x_{[j]})) f(u) du \\ &= \int_{-\infty}^{\infty} \prod_{j \neq k} F(u+d) f(u) du \\ &= \int_{-\infty}^{\infty} F^{k-1}(u+d) f(u) du = P^*, \end{aligned}$$

since $\sum p_i(\underline{x}) = 1$, so $p_{[1]}(\underline{x}) = p_{[2]}(\underline{x}) = \cdots = p_{[k-1]}(\underline{x}) = \frac{1}{k-1}(1-P^*)$ and $|s_{R^{\max}}| = i$, hence the result follows. □

Since $PP(CS|R^{\max}, \underline{x}) \geq P^*$ and it is strictly larger than P^*, once we pick more than one population, we certainly can find a better subset selection rule by simply utilizing the lower bound on $PP(CS|R^{\max}, \underline{x})$.

4.3 Some New Selection Procedures

First, let us consider the following selection procedure:

Let $\Delta x_{[i]} = x_{[k]} - x_{[k-i]}$ for $i = 1, 2, \cdots, k-1$, where $x_{[1]} \leq x_{[2]} \leq \cdots \leq x_{[k]}$ are the ordered values of x_i's. then, we compute the following $k-1$ numbers:

$$P^*_{(i)} = \int_{-\infty}^{+\infty} F^{k-1}(u + \Delta x_{[i]}) dF(u). \tag{16}$$

Since $0 \leq \Delta x_{[1]} \leq \cdots \leq \Delta x_{[k-1]}$, therefore

$$0 \leq P^*_{(1)} \leq P^*_{(2)} \leq \cdots \leq P^*_{(k-1)} (< 1).$$

Next, we compute:

$$\mathbf{Q}^*_{(i)} = P^*_{(i)} + \frac{i-1}{k-1}(1 - P^*_{(i)}). \tag{17}$$

Lemma 4.2: For values of $\Delta x_{[i]}$, where $0 \leq \Delta x_{[1]} \leq \cdots \leq \Delta x_{[k-1]}$, we have

$$0 \leq \mathbf{Q}^*_{(1)} \leq \cdot \cdot \leq \mathbf{Q}^*_{(k-1)} (< 1). \tag{18}$$

Proof: Actually, we have

$$\mathbf{Q}^*_{(i)} = 1 - \frac{k-i}{k-1}(1 - P^*_{(i)}),$$

so $\mathbf{Q}^*_{(i)}$ is increasing in i, because $k-i$ is decreasing in i and $1 - P^*_{(i)}$ is decreasing in $\Delta x_{[i]}$ (thus in i); hence the result. □

Now, we propose the following subset selection rule R_1:

For any preassigned guarantee probability $P^*(1/k < P^* < 1)$, if there exists the smallest $\mathbf{Q}^*_{(i_0)}$ which satisfies $\mathbf{Q}^*_{(i_0)} \geq P^*$, then the subset selection rule R_1 is

$$R_1: \quad \text{Select } \Pi_{(j)} \quad \text{iff}: \quad j > i_0. \tag{19}$$

The subset selection rule R_1 will take $s = \{\Pi_{(k)}, \cdots, \Pi_{(k-i_0+1)}\}$ as our selected subset.

otherwise, R_1 will select all populations.

Remark 4.3: To implement the procedure R_1, we examine the posterior probabilities at following $k-1$ stages:

Stage 1. pull all $k-2$ values of $x_{[i]},\ i=1,2,\cdots,k-2$ to the point $x_{[k-1]}$, and check if

$$\frac{k-1}{k-1}(1-P^*_{(1)}) \leq 1-P^*,$$

if the above holds, we select $s=\{\Pi_{[k]}\}$ and terminate the process, if not, we go to

Stage 2. pull all $k-2$ values of $x_{[i]},\ i\neq k-2,\ i=1,2,\cdots,k-1$ to the point $x_{[k-2]}$, check if

$$\frac{k-2}{k-1}(1-P^*_{(2)}) \leq 1-P^*,$$

if it holds, we select $s=\{\Pi_{[k]},\ \Pi_{[k-1]}\}$ and terminate the process, if not, we go to Stage 3, and so on, until we can find an i such that

$$\frac{k-i}{k-1}(1-P^*_{(i)}) \leq 1-P^*,$$

and then we select $s=\{\Pi_{[k]},\ \Pi_{[k-1]},\ \cdots,\ \Pi_{[k-i+1]}\}$; If there does not exist such an i, we select all populations.

For other subset selection rules $R_2,\cdots,R_{k-1}$, we give the following remark:

Remark 4.4: Note that in the Process of deriving the subset selection rule R_1, we divided the data into two groups, and put only one value (i.e. $x_{[k]}$) into the first group. Now we can develop it in two directions.

(a) By putting more $x_{[i]}$'s into the first group, we can actually replace $Q^*_{(i)}$ by $Q^{**}_{(i)}$ as follows:

$$Q^{**}_{(i)} = \max_{0\leq m\leq i-1}(1-\frac{k-i}{k-m-1}P^{**}_{(i,m)}),$$

where

$$P^{**}_{(i,m)} = \int_{-\infty}^{+\infty} F^{m+1}(u-(x_{[k-m]}-x_{[k-i]}))dF^{k-m-1}(u),\quad m=0,1,\cdots,i-1,$$

is the posterior probability of $p_{[1]}(\underline{x})=\cdots=p_{[k-m-1]}(\underline{x})$, when we pull $x_{[k]},\cdots,x_{[k-m+1]}$ to $x_{[k-m]}$ and $x_{[k-m-1]},\cdots,x_{[1]}$ to $x_{[k-i]}$.

When $m=0$, we have

$$P^{**}_{(i,0)} = 1-P^*_{(i)},$$

which is the value we used in the rule R_1.

(b) We can also divide the data into $3, 4, \cdots, k$ groups. Let R_2 be the rule for the case of 3 groups, $\cdots$, and R_{k-1} be the rule for the case of k groups. then in the case of k groups, the subset selection rule and the previous rule R^B are identical. Later, it will be shown that R_2 can be as good as R^B and it is easier to implement from the computational viewpoint.

4.4 Properties of Subset Selection Rule

We can easily prove the following:

Proposition 4.1: The subset selection rule R_1 is better than rule $R^{\max}$, in the sense that

(a) $PP(CS|R_1) \geq P^*$, because $PP(CS|R_1) \geq Q^*_{(i_0)}$, and

(b) $s_1 \subset s_{R^{\max}}$, because $P^*_{(i_0)} \leq P^*$.

Proposition 4.2: (a) The subset selection rule R_1 and R^B will take the same action, if $x_{[1]} = \cdots = x_{[k-1]} < x_{[k]}$, or when the subset selection rule R_1 selects all populations in its selected subset.

(b) The subset selection rule R_1, R^B and $R^{\max}$ will take the same action, if the subset selection rule R_1 selects only one population.

Proposition 4.3: The subset selection rule R_1 possesses the advantage of the rule $R^{\max}$, because the forms of the involved integration for $P^*_{(i)}$ and P^* are identical.

Remark 4.5: The selection rule R_1 is like a modified rule of $R^{\max}$, where, it like that the population associated with the largest statistic possesses the probability P^* of a correct selection, and the remaining $|s| - 1$ populations in s have the P(CS) at laest equal to $\frac{|s|-1}{k-1}(1 - P^*)$.

5 An Example for Comparsion of the Several Subset Selection Rules

A data set of exponential random numbers generated by a statistical package G6-RVP designed by H.Rubin and C.Hinkle at Purdue University was given in Gupta and Leong's paper(1979), where 9 observations for each of 5 sets of double exponential random numbers with location parameters θ_i equal to $0, 2.5, 3.4, -2.0, -0.65$ were taken.

Π_1	Π_2	Π_3	Π_4	Π_5
-3.4839	-0.9839	-0.0839	-5.4839	-4.1339
-2.6762	-0.1762	0.7238	-4.6762	-3.3262
-0.3129	2.1871	3.0871	-2.3129	-0.9629
-0.2264	2.2736	3.1736	-2.2264	-0.8764
-0.1761	2.3239	3.2239	-2.1761	-0.8261
0.1462	2.6462	3.5462	-1.8538	-0.5038
0.3033	2.8033	3.7033	-1.6967	-0.3467
1.6160	4.1160	5.0160	-0.3840	0.9660
5.6924	8.1924	9.0924	3.6924	5.0424

To see how each subset selection rule performs, let

$x_i =$ the sample mean of Π_i and $y_i =$ sample median of Π_i,

then

$$\underline{x} = (x_1, \ldots, x_5)' = (0.0980, 2.5980, 3.4980, -1.9020, -0.5520)',$$

$$\underline{y} = (y_1, \ldots, y_5)' = (-0.1761, 2.3239, 3.2239, -2.1761, -0.8261)'.$$

Hence the difference of x_i's and y_i's are $\Delta x_{32} = \Delta y_{32} = 0.90$, $\Delta x_{31} = \Delta y_{31} = 3.40$, $\Delta x_{35} = \Delta y_{35} = 4.05$, $\Delta x_{34} = \Delta y_{34} = 5.40$.

(a) Now, we have the following:

$PP(CS|R, \underline{x})$ **for** $R = R^B, R_1, R_i (i \geq 2)$

when one population is picked

	using mean	using median
R^B or $R_i (i \geq 2)$	0.9131	0.9380
$R^{\max}$ or R_1	0.7700	0.8292

where, in the case of the sample mean, the integration for R^B is

$$P^* = \int_{-\infty}^{\infty} F_9(u + 0.9) \times F_9(u + 3.4) \times F_9(u + 4.05) \times F_9(u + 5.4) dF_9(u).$$

The integration for R_2 is

$$P^* = \int_{-\infty}^{\infty} F_9(u + 0.9) \times F_9^3(u + 3.4) dF_9(u).$$

Also, the integration for R_1 or $R^{\max}$ is

$$P^* = \int_{-\infty}^{\infty} F_9^4(u + 0.9) dF_9(u),$$

where $F_9(.)$ is the cdf of the sample mean of size 9.

The same applies to the case of the sample median. Note that the rule R_2 is as good as R^B.

(b) In the case where two populations are taken, we have the probability one for all selection rules, because

$$\int_{-\infty}^{\infty} F_9^4(u + 3.4) dF_9(u) \simeq \int_{-\infty}^{\infty} G_4^4(u + 3.4) dG_4(u) \simeq 1,$$

where $G_4(.)$ is the cdf of the sample median of size 9(see Gupta and Leong (1979)).

Table 1: Table of $\{c_n\}$ and $\{c_{n,i}\}$ for $n = 2, 3, \cdots, 10$

	sample size n								
$c_{n,i}$	2	3	4	5	6	7	8	9	10
c_n	4	16	96	768	7680	92160	1290240	20643840	371589120
$c_{n,0}$	1	3	15	105	945	10395	135135	2027025	34459425
$c_{n,1}$	1	3	15	105	945	10395	135135	2027025	34459425
$c_{n,2}$		1	6	45	420	4725	62370	945945	16216200
$c_{n,3}$			1	10	105	1260	17325	270270	4729725
$c_{n,4}$				1	15	210	3150	51975	945945
$c_{n,5}$					1	21	378	6930	135135
$c_{n,6}$						1	28	630	13860
$c_{n,7}$							1	36	990
$c_{n,8}$								1	45
$c_{n,9}$									1

Table 2: Table of $\{c_n\}$ and $\{a_{n,i}\}$ for $n = 2, 3, \cdots, 10$

	sample size n								
$a_{n,i}$	2	3	4	5	6	7	8	9	10
c_n	4	16	96	768	7680	92160	1290240	20643840	371589120
$a_{n,0}$	2	8	48	384	3840	46080	645120	10321920	185794560
$a_{n,1}$	1	5	33	279	2895	35685	509985	8294895	151335135
$a_{n,2}$		1	9	87	975	12645	187425	3133935	58437855
$a_{n,3}$			1	14	185	2640	41685	729330	14073885
$a_{n,4}$				1	20	345	6090	114765	2336040
$a_{n,5}$					1	27	588	12558	278019
$a_{n,6}$						1	35	938	23814
$a_{n,7}$							1	44	1422
$a_{n,8}$								1	54
$a_{n,9}$									1

Table 3: 1) Upper $100(1-\alpha)$ Percentage Points ξ_α of Z_n(Top Entry);
2) $\Delta_\alpha = \xi^*_\alpha - \xi_\alpha$, where ξ^*_α is the Upper $100(1-\alpha)$ Percentage Points of Z^*_n(Bottom Entry).

	sample size n									
$1-\alpha$	3	5	7	9	11	13	15	17	19	21
0.750	0.6050	0.6321	0.6440	0.6507	0.6550	0.6580	0.6602	0.6619	0.6632	0.6643
	-.0825	-.1102	-.1249	-.1343	-.1409	-.1459	-.1499	-.1531	-.1557	-.1579
0.900	1.2221	1.2432	1.2532	1.2590	1.2629	1.2656	1.2676	1.2692	1.2704	1.2715
	-.0739	-.1259	-.1591	-.1823	-.1995	-.2129	-.2237	-.2326	-.2401	-.2467
0.950	1.6372	1.6385	1.6395	1.6402	1.6408	1.6412	1.6416	1.6419	1.6422	1.6424
	-.0369	-.1025	-.1484	-.1815	-.2066	-.2264	-.2426	-.2560	-.2675	-.2774
0.975	2.0284	2.0026	1.9905	1.9836	1.9794	1.9763	1.9740	1.9724	1.9711	1.9699
	.0144	-.0632	-.1210	-.1637	-.1967	-.2229	-.2443	-.2623	-.2778	-.2910
0.990	2.5214	2.4524	2.4194	2.3999	2.3871	2.3782	2.3715	2.3663	2.3621	2.3590
	.0976	.0055	-.0683	-.1236	-.1666	-.2014	-.2298	-.2539	-.2741	-.2923
0.995	2.8821	2.7759	2.7246	2.6941	2.6746	2.6599	2.6495	2.6410	2.6343	2.6294
	.1684	.0659	-.0195	-.0842	-.1355	-.1758	-.2099	-.2387	-.2625	-.2844

Table 4: Upper $100(1-\alpha)$ Percentage Points ξ_α of Z_n for even values of n

	sample size n									
$1-\alpha$	2	4	6	8	10	12	14	16	18	20
0.750	0.5731	0.6218	0.6390	0.6478	0.6531	0.6566	0.6592	0.6611	0.6626	0.6638
0.900	1.1986	1.2350	1.2489	1.2564	1.2611	1.2643	1.2667	1.2685	1.2698	1.2710
0.950	1.6359	1.6379	1.6390	1.6399	1.6405	1.6411	1.6415	1.6418	1.6420	1.6423
0.975	2.0563	2.0125	1.9955	1.9867	1.9814	1.9777	1.9751	1.9733	1.9717	1.9705
0.990	2.5958	2.4792	2.4335	2.4084	2.3929	2.3822	2.3746	2.3688	2.3642	2.3605
0.995	2.9944	2.8174	2.7466	2.7075	2.6831	2.6666	2.6544	2.6453	2.6379	2.6318

Table 5: Values of $d/\sqrt{n} = d(n, k, P^*)$ for $n, k = 2, 3, \cdots, 10$

		number of Populations k								
n	P^*	2	3	4	5	6	7	8	9	10
	0.75	1.1462	1.7849	2.1575	2.4258	2.6365	2.8104	2.9584	3.0874	3.2015
1	0.90	2.3972	3.0504	3.4336	3.7083	3.9234	4.1002	4.2503	4.3809	4.4964
	0.95	3.2716	3.9322	4.3197	4.5971	4.8138	4.9917	5.1427	5.2739	5.3899
	0.99	5.1910	5.8612	6.2549	6.5350	6.7538	6.9333	7.0853	7.2180	7.3343
	0.75	0.9580	1.3575	1.6011	1.7756	1.9110	2.0214	2.1144	2.1947	2.2653
2	0.90	1.7893	2.1966	2.4393	2.6118	2.7452	2.8539	2.9454	3.0244	3.0939
	0.95	2.3470	2.7550	2.9962	3.1670	3.2992	3.4069	3.4975	3.5758	3.6447
	0.99	3.5237	3.9287	4.1660	4.3337	4.4634	4.5696	4.6582	4.7351	4.8032
	0.75	0.7379	1.1187	1.3251	1.4666	1.5740	1.6605	1.7328	1.7949	1.8493
3	0.90	1.4421	1.7960	1.9924	2.1281	2.2316	2.3152	2.3853	2.4455	2.4983
	0.95	1.8926	2.2339	2.4250	2.5576	2.6591	2.7410	2.8098	2.8690	2.9209
	0.99	2.8096	3.1318	3.3142	3.4417	3.5394	3.6189	3.6855	3.7427	3.7932
	0.75	0.6478	0.9796	1.1575	1.2784	1.3696	1.4428	1.5037	1.5558	1.6013
4	0.90	1.2564	1.5601	1.7268	1.8414	1.9283	1.9983	2.0567	2.1068	2.1507
	0.95	1.6399	1.9298	2.0909	2.2020	2.2866	2.3548	2.4119	2.4608	2.5038
	0.99	2.4086	2.6774	2.8290	2.9344	3.0150	3.0802	3.1351	3.1820	3.2234
	0.75	0.5842	0.8821	1.0407	1.1480	1.2286	1.2931	1.3466	1.3923	1.4322
5	0.90	1.1280	1.3980	1.5454	1.6462	1.7224	1.7836	1.8346	1.8783	1.9165
	0.95	1.4673	1.7236	1.8650	1.9623	2.0362	2.0956	2.1452	2.1877	2.2249
	0.99	2.1401	2.3752	2.5071	2.5983	2.6682	2.7246	2.7719	2.8121	2.8477

Table 5 (continued)

	0.75	0.5361	0.8088	0.9532	1.0507	1.1237	1.1819	1.2301	1.2713	1.3072
6	0.90	1.0320	1.2777	1.4110	1.5022	1.5707	1.6256	1.6714	1.7106	1.7450
	0.95	1.3396	1.5718	1.6992	1.7864	1.8530	1.9058	1.9504	1.9885	2.0215
	0.99	1.9453	2.1533	2.2734	2.3555	2.4170	2.4668	2.5078	2.5430	2.5752
	0.75	0.4983	0.7514	0.8850	0.9748	1.0420	1.0955	1.1397	1.1775	1.2103
7	0.90	0.9575	1.1843	1.3071	1.3906	1.4535	1.5038	1.5457	1.5814	1.6126
	0.95	1.2408	1.4542	1.5713	1.6513	1.7119	1.7605	1.8010	1.8356	1.8658
	0.99	1.7952	1.9874	2.0951	2.1694	2.2258	2.2714	2.3095	2.3423	2.3708
	0.75	0.4675	0.7045	0.8295	0.9132	0.9758	1.0255	1.0666	1.1017	1.1321
8	0.90	0.8969	1.1086	1.2230	1.3006	1.3590	1.4057	1.4444	1.4775	1.5064
	0.95	1.1609	1.3596	1.4683	1.5426	1.5987	1.6436	1.6810	1.7130	1.7410
	0.99	1.6750	1.8530	1.9526	2.0211	2.0731	2.1152	2.1504	2.1804	2.2068
	0.75	0.4419	0.6658	0.7835	0.8623	0.9210	0.9677	1.0063	1.0391	1.0676
9	0.90	0.8468	1.0461	1.1535	1.2263	1.2811	1.3248	1.3610	1.3920	1.4189
	0.95	1.0950	1.2816	1.3835	1.4530	1.5055	1.5475	1.5825	1.6123	1.6384
	0.99	1.5764	1.7430	1.8358	1.8997	1.9482	1.9874	2.0200	2.0480	2.0726
	0.75	0.0015	0.5712	0.7179	0.8051	0.8665	0.9135	0.9516	0.9835	1.0110
10	0.90	0.7512	0.9761	1.0869	1.1593	1.2126	1.2547	1.2893	1.3188	1.3443
	0.95	1.0141	1.2070	1.3078	1.3752	1.4255	1.4655	1.4986	1.5269	1.5515
	0.99	1.4865	1.6476	1.7362	1.7968	1.8428	1.8796	1.9103	1.9365	1.9596

References

[1] Bechhofer, R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics* **25** 16-39.

[2] Gupta, S.S. (1956). On a decision rule for a problem in ranking means. Ph.D. Thesis (Mimeograph Series No. 150), Institute of Statistics, University of North Carolina, Chapl Hill, North Carolina.

[3] Gupta, S.S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics* **7** 225-245.

[4] Gupta, S.S. and Leong, Y.-K. (1979). Some results on subset selection procedures for double exponential populations. in: *Decision Information* (Tsokos, C.P. Thrall, R.M. ed.), 277-305, Academic Press, New York.

[5] Gupta, S.S. and Yang, H.M. (1985) . Bayes-P^* subset selection procedure for the best population. *Journal of statistical planning and inference* **12** 213-233.

[6] Han, S.H. (1987). Contributions to Selection and Ranking Theory with Special Reference to Logistic Populations. Ph.D. Thesis (Report Series No. 87-38), Department of Statistics, Purdue University, West Lafayette, Indiana.

[7] Weida, F.M. (1935). On certain distribution functions when the law of the universe is Poisson's first law of error. *Annals of Mathematical Statistics* **6-7** 102-110.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release, distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Technical Report #90-25C	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Purdue University 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Department of Statistics West Lafayette, IN 47907	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Office of Naval Research 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER NSF DMS-8702620 N00014-88-K-0170
8c. ADDRESS (City, State, and ZIP Code) Arlington, VA 22217-5000	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
CLASSICAL AND BAYES-P* SUBSET SELECTION PROCEDURES FOR DOUBLE EXPONENTIAL POPULATIONS

12. PERSONAL AUTHOR(S)
Shanti S. Gupta and Yuning Liao

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM ____ TO ____	May 1990	21

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Classical Maximum-Type Procedure; Posterior Probability of a Correct Selection; Subset Selection; Selected Subset Size

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The exact distribution of the sample mean from a double exponential (Laplace) model is derived. A *classical subset selection procedure* based on the sample mean for selecting the population associated with the largest location parameter of k double exponential (Laplace) distributions is studied. For the case when a non-informative prior is introduced into the problem, the relation between the classical Maximum-Type Procedure Rule R^{max} and the so-called Bayes-P* subset selection procedure rule is studied. An improved bound for the guarantee probability of a correction selection for the classical subset selection rule R^{max} that relates the rule R^{max} to the selected subset size (notice that the subset selection rule R^{max} may select all the populations) is studied and some improved rules of the type R^{max} are provided.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Shanti S. Gupta	22b. TELEPHONE (Include Area Code) 317-494-6031 / 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED